基于特征相关的偏最小二乘特征选择方法

曾青霞[a]，杜建强[a]，朱志鹏[a]，聂　斌[a]，余日跃[b]，喻　芳[a]

(江西中医药大学 a. 计算机学院; b. 药学院，南昌 330004)

摘　要：针对传统的偏最小二乘法只考虑单特征的重要性以及特征之间存在冗余和多重共线性等问题，将特征之间的统计相关性引入到传统的偏最小二乘分析中，构造了一种基于特征相关的偏最小二乘模型。首先利用特征相关度对特征进行评估预选出特征组，然后将其放入偏最小二乘模型中进行训练，评估该特征组是否可取。结合前向贪心搜索策略依次评价候选特征，并选中使目标函数最小的候选特征加入到已选特征。分别采用麻杏石甘汤君药止咳、平喘和 UCI 数据集进行分析处理，实验结果表明，该特征选择方法能较好寻找较优的特征组。

关键词：中医药信息；偏最小二乘法；特征相关；特征选择

中图分类号：TP

PLS feature selection method based on feature correlation

Zeng Qingxia[a], Du Jianqiang[a], Nie Bin[a], Yu Riyue[b], Yu Fang [A], Huang Canyi [A]

(*School of Computy Jiangxi University of Traditional Chinese Medicine*, *Nanchang* 330004, *China*)

Abstract: The traditional partial least squares method only considers the importance of single features and it exists the redundancy and multicollinearity among the features. The statistical correlation between features is involved into the traditional partial least squares analysis, and constructed the model of PLS feature selection based on feature correlation. Firstly, the feature group is pre-selected by using of the feature relevance, and then put into the partial least squares model for training to assess whether the feature group is desirable. Combining with the greedy search strategy, the candidate features are evaluated one by one, and the candidate features with the smallest objective function are added to the selected features. Respectively, using the data of the maxingshigan decoction of the monarch drug to treat the asthma or cough and UCI data sets to analyze . The experimental results show that the feature selection method can find an optimal feature group.

Key Words: TCM information ; partial least squares; feature correlation ; feature selection

0　引言

随着科学的发展，数据挖掘领域需处理的对象越来越复杂，其数据维度也在急剧增加。较高的维数容易引发“维数灾难”，随着维数的增加，计算复杂度显著提高而分类器的性能急剧下降。因此，必须对数据进行特征降维，特征降维有两种方式：特征选择和特征提取。

特征选择是指在原始特征空间中选择能让给定任务的评价准则达到最优的特征子集的过程，是模式识别、机器学习等领域中数据预处理的关键步骤之一[1-2]。其主要目的是在不显著降低分类精度的情况下，选择一个最优的特征子集，并且移除不相关或冗余的特征，使留下的特征具有更强的分辨率[3]。其中评价准则是特征选择算法中的关键步骤，包括距离度量、信息度量、依赖性度量以及一致性度量。在数据挖掘中，基于评价准则特征选择可分为 filter[4]（筛选法）、wrapper[5]（封装法）以及 embedded（嵌入式）三类。Filter 需要评价特征相关性的评分函数和阈值判别法来选择出得分最高的特征子集，Filter 训练速度快，但评估与后续学习算法的性能偏差较大。Wrapper 利用后续学习算法的训练准确率评估特征子集，偏差小计算量大，不适合大数据集。Embedded 的出现主要是为了解决 Wrapper 在处理不同数据集时，分类模型需要重构代价高等问题[6]。它将特征选择与分类模型的学习过程结合，有着高效的时空性能及较好的分类精度。

偏最小二乘法(partial least square, PLS)在自变量间存在较高相关性时，提出了一种多因变量对多因变量的回归建模方法，可以有效地解决多重共线性问题[7]. 基于这种优势，李建更等人[8]提出了基于逐步提取偏最小二乘主成分的特征选择方法，通过重复利用偏最小二乘提取主成分并选择权重较大的基因；李

作者简介：曾青霞（1995-），女，江西九江人，硕士研究生，主要研究方向为医药数据挖掘及机器学习；杜建强（1968-），男（通信作者），江西南昌人，教授，博士，主要研究方向为医药信息与数据挖掘；朱志鹏（1990-），男，湖北咸宁人，硕士研究生，主要研究方向为机器学习及医药数据挖掘；聂斌（1972-），男，江西吉安人，硕士，主要研究方向为中医药信息及数据挖掘；余日跃（1959-），男，教授，主要研究方向为中药复方；喻芳（1992-），女，湖北武人，硕士研究生，主要研究方向为医药数据挖掘及机器学习.

胜等人[9]提出了改进的量子遗传偏最小二乘特征选择方法，该算法通过赋予种群初始值并设计了一种新的适应度函数，结合偏最小二乘法进行特征选择；Nguyen 等人[10]以偏最小二乘算法作为特征降维方法，采用线性判别分析（logistic discrimination, LD）和二次线性判别分析（quadratic discrimination analysis, QDA）算法构建分类器，用于对数据的分类。

因此，本文提出了一种基于特征相关的偏最小二乘特征选择方法。利用特征相关度对特征进行评估预选出特征子集，然后将其放入偏最小二乘模型中进行训练，评估该特征子集是否可取。结合前向贪心搜索策略依次评价候选特征，并选中使目标函数最小的候选特征加入到已选特征。该特征选择方法不仅具备训练速度快、局部最优等特点，同时还弥补了 Wrapper 不适合大数据集、计算量大等缺点，从而找出较优的特征子集。

1　基于相关性的特征选择

Hall[11]于 1999 年提出基于相关性的特征选择(correlation-based feature selection, CFS)方法。CFS 方法是一种典型的 filter 式特征选择方法，它启发式地对单一特征对应于每个分类的作用来进行评价，从而得出最终的特征子集。

1.1　特征估计

CFS 估计特征子集并对特征子集而不是单个特征进行排序。其核心是采用启发式的方式来评估特征子集的价值。CFS 通过计算特征之间的相关性以及特征与类标之间的相关性来实现特征的选择，其目的是使被选中的特征之间彼此尽可能不相关，而与类标之间高度相关。CFS 的启发式方程为

$$Merit_s = \frac{kr_{cf}}{\sqrt{k+k(k-1)r_{ff}}} \tag{1}$$

其中：$Merit_s$ 表示包含 k 个特征的特征子集 S 的‘merit’（类别区分能力），r_{cf} 表示类别 c 与特征 f ($f \in S$)的平均相关系数，r_{ff} 是特征 f 之间的平均相关系数。r 为 Pearson 相关系数，所有的变量需要标准化。分子部分表示特征子集 S 的类预测能力；分母表示特征子集 S 中特征的冗余程度。因此分子越大表示特征子集 S 的类预测能力越强，分母越小表示该特征子集的冗余越小。

特征选择就是选择一组特征构成特征子集，该子集与类别高度相关，但是子集中的特征之间高度不相关。由此可见 $Merit_s$ 的值越大，当前特征子集 S 对于分类的贡献越大，是优良的特征子集。

但在 CFS 中，特征必须是离散的随机变量，而且是通过条件熵和互信息的计算方式来对特征之间和特征与类标之间进行评价。因此针对数据是连续性的随机变量时就难以处理，基于此，针对数据是连续性随机变量时可通过 Pearson 相关系数[12]来计算特征之间的相关性以及特征与类标之间的相关性。相关系数的绝对值越大，则相关性越强；相关系数越接近于 0，则相关度越弱。

1.2　搜索特征子集空间

CFS 首先从训练集中计算特征与类和特征与特征相关矩阵，然后用前向选择搜索策略(forward selection search strategy，FS)搜索特征子集空间，也可使用其他的搜索方法，包括最佳优先搜索（best first search，BFS）、后向消除(backward elimination，BE)。前向选择刚开始没有特征，然后贪心地增加一个特征直到没有合适的特征加入。后向消除开始有全部特征，然后每一次贪心地去除一个特征直到估计值不再降低。最佳优先搜索和其他两种搜索方法差不多。可以开始于空集或全集，以空集 M 为例，开始时没有特征选择，并产生了所有可能的单个特征；计算特征的估计值（由 $Merit_s$ 值表示），并选择 $Merit_s$ 值最大的一个特征进入 M ，然后选择第二个拥有最大的 $Merit_s$ 值的特征进入 M ，如果这两个特征的 $Merit_s$ 值小于原来的 $Merit_s$ 值，则去除这个第二个最大的 $Merit_s$ 值的特征，然后在进行下一个，依次递归，找出使 merit 最大的特征组合。

2　基于特征相关的偏最小二乘特征选择（PLS feature selection based on feature correlation，PLSCF）

偏最小二乘回归[13](PLS)是一种新型的多元统计分析方法，与传统的最小二乘回归不同，偏最小二乘回归研究的是多因变量对多自变量的回归建模。特别是当变量存在多重相关性或样本点数据少于变量个数的时候,采用偏最小二乘回归模型更为有效。

2.1　偏最小二乘回归建模思想

存在自变量集合 $X = (x_1, x_2, x_3, \ldots, x_n)$ 和因变量集合 $Y = (y_1, y_2, y_3, \ldots, y_m)$，为了能最好地概括原数据信息的综合变量,在 X 中提取第一个成分 t_1，使得方差 $Var(t_1) \to max$ 。在 Y 中提取第一个成分 u_1 ，使得方差 $Var(u_1) \to max$ ，并使得相关性 $r(t_1, u_1) \to max$ 。然后将 t_1 和 u_1 进行多元线性回归，得到残差向量，用同样的方法依次迭代。用交叉有效性确定偏最小二乘回归中所需要提取的主成分个数，停止迭代，建立偏最小二乘回归模型。

2.2　基于 PLSCF 的前向选择搜索策略算法

以 CFS 度量相应特征子集的类间区分能力和 PLS 回归模型的残差平方和（sum of squares for error, SSE)作为选择相应特征子集的评价指标，称这种方法为 PLSCF 评价准则。而搜索策略采用前向选择。

该算法将 PLSCF 特征评价准则与前向选择搜索策略结合。首先加入最具有类间区分能力的一个特征，然后迭代加入与已选择特征组合最具有类间区分能力的相应特征，之后浮动部分依据加入特征之后的特征子集对应 PLS 模型的残差平方和判定加入的特征是否保留。若当前特征子集训练所得 PLS 的 SSE 下降，则保留加入的特征，否则删除加入的特征。依次重复实验，直到所有特征都被测试过。最后留在特征子集中的特征构成被选中的最佳特征子集。算法伪代码描述见算法 1。

算法 1 基于 PLSCF 的 FS 混合特征选择算法

输入：当前训练集和测试集

输出：特征子集 C

Setp1 将数据进行预处理

Step2 特征估计

设 $S=\{f_i \mid i=1,2,\ldots,m\}$ 为全部特征构成的集合，C 为被选择特征构成的子集，初始为空集，即 $C=\varnothing$

while $S \neq \varnothing$ DO

根据 $Merit_s = \frac{kr_{cf}}{\sqrt{k+k(k-1)r_{ff}}}$，在训练集上计算每个特征的区分能力

选择最重要的特征 $f_{\max} = \max\{F_i, i=1,2,\ldots,m\}$

Step3 使用前向选择搜索策略评价候选特征值

令 $S = S - \{f_{\max}\}$

令 $C = C \cup \{f_{\max}\}$

使用 C 中的特征训练 PLS，得到一个 PLS 预测模型

记录该模型对训练集和测试集的残差平方和 *SSETrain* 和 *SSETest*

if *SSETrain* ≥ *preSSETrain* then

$C = C - \{selected \quad f_{\max}\}$

Step4 end

3 实验结果及分析

本文的实验数据主要来源于江西中医药大学重点实验室的麻杏石甘汤止咳数据(MXZK)、平喘数据(MXPC)和 UCI 数据集的 Air Quality、CASP、Slump、Housing 和 CCPP_Folds5x2_pp[14]。

3.1 实验数据说明

麻杏石甘汤咳喘数据和 UCI 数据集 Air Quality、CASP、Slump、Housing、ENB2012_data、CBM Dataset、CCPP_Folds5x2_pp 的基本信息如表 1 所示。

表 1 数据集信息

数据集	自变量个数	因变量个数	样本数
MXZK	5	1	62
MXPC	5	1	46
AQ	11	1	9357
CASP	9	1	45730
Slump	7	3	103
Housing	13	1	506
CCPP	4	1	9567

3.2 实验结果及分析

为验证提出的 PLSCF 的特征选择方法的可行性和有效性，将七个数据集分别采用支持向量机(SVM)、基于相关性的特征选择(CFS)以及基于特征相关的偏最小二乘特征选择(PLSCF)进行实验比较，并采用前向选择搜索策略搜索子集。将数据按照 7:3 的比例随机划分，70%构建学习训练集，30%做测试。为了得到具有统计意义的实验结果，在实验的具体过程中，通过调整模型参数使得模型达到最优，且在同一学习训练集的水平下对两种算法效果进行比较。分别考察训练集残差平方和(sum of squares for error of train, SSETrain)和测试集残差平方和(sum of squares for error of test, SSETest)。实验结果如表 2 所示。

根据表 2 的实验结果可知，在以上七组数据集上，用 SVM 算法与 CFS 算法并结合 FC 搜索策略进行特征选择所得出的训练集和测试集的残差平方和相差不大，说明两者对于以上类型数据进行特征选择的效果差不多。例如，在 CCPP 数据上，两种算法的训练集和测试集的残差平方和分别为 100.3872 和 112.4920、4.2302 和 6.5398。在而对于提出的 PLSCF 方法并结合 FC 搜索策略进行特征选择所得的测试集和训练集的残差平方和，相比较前两种算法有着明显的降低。例如：在数据集 AQ 上，三种算法的测试集和训练集残差平方和分别为 4.6118、3.7188、0.2328 和 0.0385、0.0894、0.0106。在 CASP、Housing 以及 CCPP 数据集上也是如此。而在数据集 MXZK、MXPC 以及 Slump 上，三种算法得出的训练集和测试集的残差平方和相差并不明显。其中，在 Slump 数据集上，CFS 算法的训练集的残差平方和上小于 PLSCF 算法，分别为 0.3049 和 0.3091，但测试集的残差平方和大于 PLSCF 算法，分别为 0.0312 和 0.03041，这是因为不同的数据有着不同的实验效果且所选择的特征子集并非全局最优，只能是较优。SVM 和 CFS 的运行时间普遍比 PLSCF 少，这是因为 PLSCF 特征选择算法在每次特征评价时需要用到 PLS 算法，增加了程序的运行时间。

表 2 实验结果比较

数据集	SSETrain			SSETest			runtime (ms)		
	SVM	CFS	PLSCF	SVM	CFS	PLSCF	SVM	CFS	PLSCF
MXZK	0.5321	0.6497	0.4275	12.4512	13.7281	11.9978	24	23	54
MXPC	2.4712	3.5602	1.4352	19.4212	17.5431	15.3214	40	34	59
AQ	4.6118	3.7188	0.2328	0.0385	0.0894	0.0106	304	215	1041
CASP	2378.6302	3464.840	2576.299	300.1425	308.7589	224.3983	342	452	2205
Slump	0.4218	0.3049	0.3091	0.0216	0.0312	0.03041	43	45	72
Housing	17.1057	14.0089	6.2314	0.5402	0.6527	0.3365	65	75	124
CCPP	100.3872	112.4920	32.7869	4.2302	6.5398	2.0724	1204	1025	2418

为了更直观地显示实验结果，分别绘制图 1 和 2 以体现训练集残差平方和和测试集残差平方和的波动情况。由于各个数据集的训练集和测试集的数量级不同，为了方便比较各数据集在不同算法上的测试集的残差平方和与训练集的残差平方和的波动情况，将它们统一数据中心化到[0,1]。数据中心化采用公式：

$$x_{ij}^{*} = \frac{x_{ij}}{\max_j \{x_{ij}\}} \tag{2}$$

分别将训练集和测试集的残差平方和进行中心化处理，根据该公式，使得图形在一个数量级别上方便进行比较，绘制出

图 1 和 2。

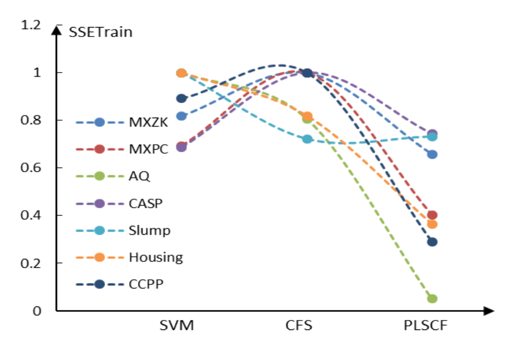


图 1 七组实验数据下各方法 SSETrain 比较

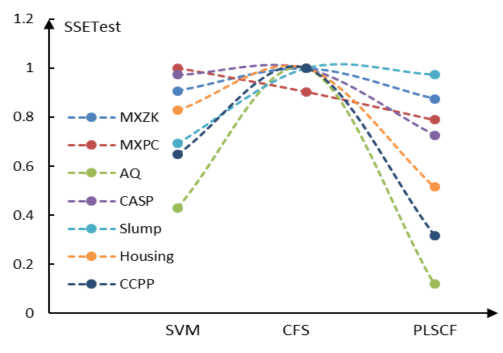


图 2 七组实验数据下各方法 SSETest 比较

由图 1 和 2 可以更加直观地看出，在 Slump 数据集上，PLSCF 训练集的残差平方和大于 CFS 小于 SVM，说明 PLSCF 算法相比较 CFS 而言不甚理想。测试集的残差平方和远大于 SVM，说明 PLSCF 算法相比较 SVM 效果略差。除 Slump 数据集效果不太明显外，PLSCF 在各项指标中效果都有明显提升，效果比 SVM 和 CFS 算法都要好。这是因为实验数据的不同，特征选择子集存在一定的随机性，不能保证全局最优，只能是较优。以上数据说明，由于实验数据选取的不同，算法效果也存在着一定的差异性。除此之外，使用基于特征相关的偏最小二乘特征选择方法的训练集和测试集的残差平方和在其他实验数据中均呈现明显的下降趋势。

综上所述，在以上七组实验中，PLSCF 方法明显优于 SVM 和 CFS，但对于个别数据集，效果不显著，是因为采用基于特征相关的偏最小二乘的评价准则的方法选出的较优特征子集虽然具有更好的代表性，但是依据不同的实验数据有着不同的效果，这也表明所选择的特征子集可能不是全局最优的，只能是较优的。

4 结束语

本文针对传统的偏最小二乘法只考虑单特征的重要性以及特征之间存在冗余和多重共线性等问题，将特征之间的统计相关性引入到传统的偏最小二乘分析中，提出了一种基于特征相关的偏最小二乘特征选择方法。充分利用了特征子集区分度的评价准则，并结合了能在样本量少的情况下依旧可以回归建模以及最大化自变量和因变量之间的关系的 PLS，充分发挥了算法各自本身的优点。通过在中医药数据以及 UCI 数据集的实验比较，与 SVM 和 CFS 算法进行对比分析，基于特征相关的偏最小二乘特征选择方法选出的特征子集具有更好的代表性。

参考文献：

[1] Lin Yaojin, Li Jinjin, et al. Feature selection via neighborhood multi-granulation fusion [J]. Knowledge-Based Systems, 2014, 67 (3): 162-168.

[2] Zhang Ce, Arun K, Christopherré, Materialization optimizations for feature selection workloads [J]. ACM Trans on Database Systems, 2016, 41 (1): 2.

[3] 曹晋，张莉，李凡长. 一种基于支持向量数据描述的特征选择算法 [J]. 智能系统学报, 2015 (2): 215-220.

[4] Zhu Z, Ong Y S, Dash M. Wrapper-filter feature selection algorithm using a memetic framework [J]. IEEE Trans on Systems Man & Cybernetics Part B: Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 2007, 37 (1): 70-76.

[5] Zhang X. A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine [J]. Geographical Research, 2008, 37 (3): 71471J-71471J-9.

[6] 王翔，胡学钢. 高维小样本分类问题中特征选择研究综述 [J]. 计算机应用, 2017, 37 (9): 2433-2438.

[7] 尚志刚，董永慧，李蒙蒙，等. 基于偏最小二乘回归的鲁棒性特征选择与分类算法 [J]. 计算机应用研究, 2017, 37 (3): 871-875.

[8] 李建更，耿涛，阮晓钢. 基于逐步提取偏最小二乘主成分的特征选择方法 [J]. 生物学杂志, 2010, 27 (4): 85-87.

[9] 李胜，张培林，李兵，等. 改进的量子遗传偏最小二乘特征选择方法应用 [J]. 计算机工程与应用, 2017, 53 (3): 242-252.

[10] Nguyen D V, Rocke D M. On partial least squares dimension reduction for microarray-based classification: a simulation study [J]. Computional Statistics&Data Analysis, 2004, 46 (2): 407-425.

[11] Hall M A. Correlation-based feature selection for discrete and numeric class machine learning [C]// Proc of the 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2000, 359-366.

[12] Saad Z S, Glen D R, Gang C, et al. A new method for improving functional-to-structural MRI alignment using local Pearson correlation [J]. Neuroimage, 2009, 44 (3): 839-848.

[13] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics [J]. Chemometrics & Intelligent Laboratory Systems, 2001, 58 (2): 109-130.

[14] UCI machine learning repository [EB/OL]. [2016-07-18]. http://archive. ics. uci. edu/ml/.